Journal of Librarianship and Scholarly Communication

Volume 1 | Issue 1 eP1035

The Anatomy of a Data Citation: Discovery, Reuse, and Credit

Hailey Mooney, Mark P. Newton

Mooney, H, Newton, MP. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication* 1(1):eP1035. http://dx.doi.org/10.7710/2162-3309.1035

© 2012 by the author(s). This open access article is distributed under a Creative Commons Attribution License, which allows unrestricted use, distribution, and reproduction in any medium, providing the original author and source are credited.

JLSC is a quarterly journal sponsored and published by Pacific University Library | ISSN 2162-3309 | http://jlsc-pub.org

ISSN 2162-3309 JLSC

RESEARCH

The Anatomy of a Data Citation: Discovery, Reuse, and Credit

Hailey Mooney *Data Services and Reference Librarian, Michigan State University*

Mark P. Newton *Production Manager, Center for Digital Research and Scholarship, Columbia University*

Abstract

INTRODUCTION Data citation should be a necessary corollary of data publication and reuse. Many researchers are reluctant to share their data, yet they are increasingly encouraged to do just that. Reward structures must be in place to encourage data publication, and citation is the appropriate tool for scholarly acknowledgment. Data citation also allows for the identification, retrieval, replication, and verification of data underlying published studies. **METHODS** This study examines author behavior and sources of instruction in disciplinary and cultural norms for writing style and citation via a content analysis of journal articles, author instructions, style manuals, and data publishers. Instances of data citation are benchmarked against a Data Citation Adequacy Index. **RESULTS** Roughly half of journals point toward a style manual that addresses data citation, but the majority of journal articles failed to include an adequate citation to data used in secondary analysis studies. **DISCUSSION** Full citation of data is not currently a normative behavior in scholarly writing. Multiplicity of data types and lack of awareness regarding existing standards contribute to the problem. **CONCLUSION** Citations for data must be promoted as an essential component of data publication, sharing, and reuse. Despite confounding factors, librarians and information professionals are well-positioned and should persist in advancing data citation as a normative practice across domains. Doing so promotes a value proposition for data sharing and secondary research broadly, thereby accelerating the pace of scientific research.

Implications for Practice:

- Promotion of data citation will foster a scholarly communication system that allows for identification, retrieval, and attribution of research data.
- Repositories publishing data should include appropriate metadata and mandate citations as a condition of reuse.
- Identifying available data via the published literature will always be a problematic reference strategy until consistent citation provides a standard retrieval mechanism.
- Normalizing expectations for dataset citation will incentivize data sharing and promote secondary research, improving the pace and quality of scholarly exchange.

© 2012 Mooney & Newton. This open access article is distributed under a Creative Commons Attribution 3.0 Unported License, which allows unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

INTRODUCTION

The institutions of science and academia are made up of communities and individuals working towards a common goal: advancing knowledge. Yet altruism is only one component of individual motivation within the scientific enterprise. Receiving credit, advancing careers, and securing grants and tenure are all necessary to the success of individual researchers and their supporting institutions.

Knowledge advances by building upon the work of those who came before us and data itself are the building blocks of knowledge. We are in the age of data-intensive science and the movement towards sharing research data is growing in momentum. The combined weight of common and individual needs are pushing the data sharing movement to reveal the inadequacies of the infrastructure that support the scientific institution; both physical technological needs and cultural practices. One such inadequacy is the lack of acknowledgement and credit given to authors of published research data. The data sharing movement needs citations for data to become common practice.

Given the need for the normative inclusion of citations for data in the scholarly literature and the perception that there is still significant work to be done for this to happen, our study seeks to uncover the actual state of practice and instruction in this area. This study will characterize the current state of data citation standards and instruction, and the current state of data citation practice across the breadth of academic research, through a content analysis of journal articles, style manuals, and journal guidelines. These interconnected facets of the scholarly communication system are bench-marked against a Data Citation Adequacy Index in order to examine the efficacy of current practices.

LITERATURE REVIEW

Advocates for the citation of research data have been raising their voices for decades in concert with the movement toward increased data publication and sharing. The 1960s ushered in the development and organization of social science data archives (Heim, 1987) and by 1979 the first set of formal guidelines for the citation of research datasets was published (Dodd, 1979). By the 1980s the "reward dilemma" was seen as easily identifiable in a research culture where sharing data had "no place on the curriculum vita," leading to identification of the obvious solution to strengthen rewards in part by the improvement of citation practices (Clubb, Austin, Geda, & Traugott, 1985, p. 58). The National Research Council's Committee on National Statistics made an explicit recommendation that "journals should require full credit and appropriate citation to original data collections in reports based on secondary analysis" in order to encourage data sharing (Fienberg, Martin, & Straf, 1985, p.31).

Even when citation has not been mentioned explicitly, the need for incentives to encourage data sharing is widely acknowledged. A survey of data sharing attitudes conducted in 1985 revealed that although most scientists agree that data sharing is a desirable practice in theory, the fear of receiving no credit and losing funding or publishing opportunities is a serious deterrent to actual practice (Ceci, 1988). Stanley & Stanley (1988) echo the fear that sharing data could result in someone else publishing with no reward given to the sharer since there is no system of acknowledgement, and Baron (1988) debates the efficacy of data sharing policies without a way to "[tabulate] research 'assists' as a routine part of the academic scorecard" (p. viii). Biologist Joshua Lederberg also observed the lack of a system in place for giving credit, noting that "some fairly famous cell lines were generated by obscure people," giving rise to the observation that "if people were rewarded for contributing to data banks...it would enhance the 'scientific ethos'" (Marshall, 1990, p. 957).

Current observations remain essentially the same. Among the reasons researchers are reluctant to share data, priority concerns (regarding credit and intellectual property rights) and lack of reward continue to be issues (see for example, Borgman, 2007, pp.196-201; LeClere, 2010). The need for citation is consistently mentioned in recent investigations of scientists' data sharing practices and perceptions. For example, in a survey of computational scientists (Stodden, 2010) the fear of use without proper citation was one of the top reasons offered for not sharing data, after the time required to prepare data for release. Stodden notes that this reveals an "incentive misalignment in the reward structure for scientific research" as "many aspects of research are tedious and time consuming, yet they get done when the expectations and reward structures are in place" (p. 21). The issue of improper citation was cited as a misuse of shared research data and a clear

deterrent to sharing in the Data Curation Profiles project interviews (Cragin, Palmer, Carlson, & Witt, 2010). Another series of interviews conducted by the Research Information Network revealed that scientists would like to see "standard, workable mechanisms for citing datasets" (Swan & Brown, 2008, p. 26) as an incentive for publishing their research data. The report also suggests the "data paper" as a feasible way for journals to provide a formal way of citing datasets (p. 12), an idea recently popularized by the California Digital Library (see Kunze et al., 2011).

Tenopir et al. (2011) surveyed scientists' data sharing practices and perceptions and found that "along with the ability to place some restrictions on sharing for some of their data, the most important condition for sharing data is to receive proper citation credit when others use their data" (p. 9). Furthermore, 93 percent of respondents thought that a fair condition of data reuse would be to provide a formal acknowledgement of the data provider and 95 percent agreed that a fair condition of data reuse would be to provide a formal citation in works that make use of the data (p. 10). Despite this consensus on the provision of a citation as a condition of data reuse, researchers are not consistently citing datasets in instances of secondary analysis in the published literature.

Lack of consistent data citation is evidenced by anecdotal observations and by studies of data reuse. In a content analysis of papers based on secondary analysis of the General Social Survey, Sieber and Trumbo (1995) found that just 19 percent of authors included the name of the survey within their references. Another content analysis by Mooney (2011) examined articles from the ICPSR Bibliography of Data Related Literature and reported that 29 percent of authors provided a complete data citation in the reference list. In an analysis of papers using the Moderate Resolution Imaging Spectroradiometer snow cover data from the National Snow and Ice Data Center, only a small fraction cite the dataset formally, substantiating the observation that "authors rarely cite data formally in journal articles and often lack guidance on how data should be cited" (Parsons, Duerr, & Minster, 2010).

The issue of guidance for the citation of data is key. Authors look to style manuals, content providers, and journals for instruction in proper bibliographic formatting. Style manuals in particular reflect disciplinary discourse norms (Hagge, 1997) and make the important connection between style and scholarly integrity (Walker & Taylor, 2006). The major style guides, the de facto arbiters of what is citable and how to cite, have yet to comprehensively address the growing needs of scientists and humanists alike to cite the data that underlies their work (Newton, Mooney & Witt, 2010).

METHODS

An author's decision to cite (or not cite) data is influenced by standards which are normalized and codified by style manuals, journal policies, and data providers. These three elements: citation styles, publishing standards, and author behavior, are all part of an interconnected system that form the basis of the scholarly journal publishing context within which data citations exist. Given the trio of facets identified for investigation, there are three distinct questions this study seeks to answer:

Part 1. *What constitutes an adequate data citation: what are the key elements needed in a data citation?*

The first component is an examination of standards indifferent of style. That is, a look at the common elements of citations without regard for particular formatting conventions. We assembled an aggregate view of existing standards in pursuit of a new tool suited for the context of the present study, a new Data Citation Adequacy Index (DCAI). We began with a survey of the history of published data citation standards in order to examine the ways in which the proposed rubrics for citing data have evolved with other standards, technologies related to the communication of digital data, and general trends in scholarship.

Part 2. *How do actual citations in journal articles measure up against instruction and best practice?*

To answer this question we conducted a content analysis of author data citation behavior. We looked at actual examples of authors using data in secondary analyses and utilized the DCAI to evaluate the efficacy and prevalence of references to the source data.

Part 3. *What kind of instructions for the citation of data are provided to authors?*

Style manuals, journal guidelines, and the sources of data

themselves are all in a position to provide instruction to authors on the proper citation of data. Each of these sources is characterized, with comparison made back to the DCAI as our standard benchmark.

PART 1: THE DATA CITATION ADEQUACY INDEX

To build a framework around which a study of data citation practices might be undertaken, the researchers created a new rubric, the Data Citation Adequacy Index (DCAI), to assign scores corresponding to the completeness of the data citation given in a particular journal article. Importantly, the DCAI is therefore neither intended to suggest a prescriptive approach to data citation nor to reflect the researchers' suggestion of the ideal data citation. This rubric is modified from the seminal work of Sieber & Trumbo (1995) in their analysis of the data citation practices of authors using the General Social Survey, but has been expanded to include the consideration of additional data citation standards in the selection of citation elements. Article scores are assigned based on the inclusion of key data citation elements and their location within the paper in descending rank order from three to zero as follows: references, notes, body text, or not present. The index applies a weight to the relative importance of each citation element: author, title, date, publisher, and URL are all afforded a weight value of two based on their universal presence among the various standards and essentialness to the functions of acknowledgment and retrieval. A weight value of one is given to the persistent identifier (a newer element) and the material designator, which is increasingly anachronistic given the new realities of networked, online retrieval. The utility and functioning of the DCAI is best understood in practice and is further explained in its application to the analysis of author behavior.

The elements of the DCAI were selected through consultation of the five data citation standards identified for inclusion (Starr et al, 2011; Green, 2009; Altman & King, 2007; Sieber & Trumbo, 1995; Dodd, 1979), alongside the citation formats for databases (the closest approximation of datasets) in national and international bibliographic standards (National Information Standards Organization & American National Standards Institute, 2005; International Organization for Standardization, 2010). These standards all provide prescriptive recommendations for the citation of data published across a span of more than three decades within specific contexts. This variety strengthens the DCAI by illustrating the strong commonality of data citation elements across time and type, but also introduces a confounding factor of disparate data types.

As a confounding factor, data heterogeneity figured largely throughout the study, as it was frequently clear that multiple reference works and report authors were using slight variations among the notions of data, dataset, numeric data, digital data, and so forth. For clarity, therefore, the project demanded a context-appropriate definition in answer to the question: What are data? For reference, each of the component references used to generate the DCAI uses a distinct definition (Figure 1, following page). To overcome the problems inherent in definition and data heterogeneity then, the researchers elected to formalize a working definition of data for the purposes of the study: one that speaks to the elemental functions of a data citation, namely to identify, acknowledge, and retrieve a data source.

> *Digital Research Data:* Any primary source in electronic format that is subject to (secondary) analysis.

This definition deliberately limits the frame of the discussion in the following ways:

- *Digital:* the researchers wanted only to understand the practices of data citation as now effected by our digital research environment.
- *Primary Source / Secondary Analysis:* By setting up the dichotomy between primary source materials used in secondary analysis, we look forward to the ways in which we expected to select articles for the evaluation. Further, we focus the discussion around attribution in data sharing, which is premier among the factors influencing researchers' decision to make source data available.

Selecting the elements

With a working definition in hand, and a set of reference data citation articles and reports as references, we began the work of developing the citation adequacy index by selecting a base set of citation elements. The most frequently identified fields across these external references were then adopted into our scoring system

Figure 1. Extant Data Citation Standards and Data Defined

Standard	Data Type	Example Citation
DataCite (Starr et al, 2011)	**Scientific research data on the Internet** "Please note that in this document, the resource that is being described can be of any kind, but it is typically a dataset. We use the term 'dataset' in its broadest sense. We mean by it to include not only numerical data, but any other research data outputs." (p. 5)	Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. doi: 10.1594/WDCC/dphase mpeps. http://dx.doi.org/10.1594/WDCC/dphase_mpeps
OECD (Green, 2009)	**OECD Data** "Taking these requirements into account, OECD is proposing to implement a metadata standard for publishing datasets, collections of datasets and individual tables." (p. 10)	OECD (2008), "Social Expenditures aggregates", OECD Social Expenditure Statistics (database). doi: 10.1787/000530172303 http://dx.doi.org/10.1787/000530172303(Accessed on 21 December 2008)
Dataverse Network (Altman & King, 2007)	**Quantitative data** "…no special restrictions on what constitutes a quantitative data set, a definition may be useful: A quantitative data set represents a systematic compilation of measurements intended to be machine readable. The measurements may be the result of scientific research or information produced by governments or others for any purpose, so long as it is systematically organized and described." (2 Quantitative data, para. 1)	Micah Altman; Karin MacDonald; Michael P. MacDonald, 2005, "Computer Use in Redistricting", hdl:1902.1/AMXGCNKCLU UNF:3:J0PkMygLPflyT1E/8x)/EA== http://id.thedata.org/hdl%3A1902.1%2FAMXGCNKCLU
Computing Index of Citation Adequacy (Sieber & Trumbo, 1995)	**Research data (GSS data sets)** Analysis of citations to General Social Survey data (machine readable numeric dataset)	Davis, James Allan and Smith, Tom.: General Social Surveys, 1972-1979, [machine-readable data files]. NORC ed. Chicago, National Opinion Research Center, Producer 1989; Storrs, CT; The Roper Center for Public Opinion Research, University of Connecticut, distributor. 1 data file (24,893 logical records) and 1 code book.
IASSIST Classification Action Group (Dodd, 1979)	**Social science numeric data** "Social science numeric data files make up a substantial body of information known in the generic sense as machine readable data files (MRDF)…" (p. 77)	n/a
Data Citation Adequacy Index (Mooney & Newton, 2012)	**Digital research data** Any primary source in electronic format that is subject to (secondary) analysis.	n/a

Figure 2. Developing the DCAI

This table shows a matrix of the standards used in the creation of the DCAI establishing the common elements between them. The far right column indicates the elements in the DCAI along with their weight score.

	Standards							
Citation Elements	DataCite	OECD	Dataverse Network	Computing Index of Citation Adequacy	IASSIST Classification Action Group	ANSI/NISO Z39.29	ISO 690-2010	Data Citation Adequacy Index
Author	x	x	x	x	x	x	x	2
Title	x	x	x	x	x	x	x	2
Date (of publication)	x	x	x	x	x	x	x	2
Publisher	x		x	x	x	x	x	2
Location				x	x	x	x	
Funder				x	x			
Material designator	x	x	x	x	x	x	x	1
Notes				x	x			
Edition	x		x		x			
URL		x	x			x	x	2
Persistent Identifier	x	x	x			x	x	1
Accessed date		x				x	x	
Parent/series		x			x		x	
Study/Accession number								

(Figure 2). The elements *author, title, date,* and *publisher* were common across all of the resources and so were incorporated into the DCAI with the strong weight of two. *Date* was closely scoped to mean date of data publication not date of data collection, which was a narrow distinction later in the coding session. *Material designator* was also present across them, but we felt it was an element type with limited utility for the present analysis, and so weighted it with the standard value of one. Neither Sieber & Trumbo nor Dodd made reference to electronic location and identification systems, a product of the early days in which these reports were issued. Today, leading initiatives such as DataCite are predicated on the notion that digital persistence is a critical component in data citation and retrieval. We therefore adopted the following system to weight citations with online retrieval locations: data citations with a URL were weighted with a two, while those that went the extra step to ensure that URL adhered to a specific persistent address schema were given another point weighted at one. For the purposes of this study, persistent element was limited to a few formally published schemas: DOI, Handle, PURL, and ARK.

PART 2: AUTHOR BEHAVIOR

Methods

We applied the DCAI rubric to scholarly research articles that used digital research data across many disciplines while maintaining a consistent approach to sample selection. To achieve this, we elected to sample the WilsonWeb Humanities, Social Sciences, and Science databases independently with a common query. These databases were not configured for large, bulk, or record download at the time we attempted to draw our sample, and so we negotiated with Wilson directly for a larger-than-normal allocation of search results for the purposes of this study.

The query devised to apply across the databases was as follows:

> (data OR dataset) <in> Abstract AND ("data bank" OR repository OR archive OR study OR studies OR empirical OR research OR obtain* OR retriev* OR use* OR analy*) <in> Abstract AND Feature Article <in> ARTICLE_TYPE AND Date: between 2010 and 2010 AND Limited to: PEER_REVIEWED

This search query defines our article sample in the following manner:

1. Search limited to uses of the word *data* or *dataset* in the abstract explicitly. Appearance of the word data in this context was believed to be a heavy predictor that the research itself would be data-driven, regardless of disciplinary origin.
2. Search limited to variations on the phrases *data bank, repository, archive, studies, empirical, research, obtain, retrieval, use,* and *analysis* because these were believed to be heavy predictors (when combined with the first phrase of the AND) of evidence of secondary data analysis and digital research data specifically.
3. Peer-reviewed research articles only published in the 2010 calendar year.

The three databases yielded significantly different numbers of records. To increase the random nature of the sample, we examined only every tenth record in the static result set. Because the Humanities index result set was significantly smaller, we also draw again every tenth record beginning and 1, 11, 21 and so forth to ensure comparable numbers across the three indexes. Initially, 25 records from each index were drawn. After final analysis, having discarded records for reports not pertaining to digital research data, we analyzed 22 records from the Science index, 20 from the Humanities index, and 23 from the Social Science index.

Each article in the sample was coded twice. Approximately half of each sample was first coded by one of this paper's authors and then the other author was assigned the secondary role of recoding for consistency and discussion. Having determined that the article met the minimum criteria, the coder identified the primary digital research data dataset (or selected one in the presence of several). The article would then be combed for references to that dataset, starting with the references, but then checking acknowledgments, notes fields, supplementary materials, the data section of the body text itself, and then the remaining body text if no other reference could be located. If the data were referenced in multiple parts of the paper, then the DCAI elements in the highest-weighted locations were calculated separately from those other elements located elsewhere in the document. The authors made notes to each other of any outstanding or pertinent information that would assist in the coding of a particular article, especially when attempting to determine whether the reference to some given digital research data was a formal citation to the actual source material or rather (as was frequently the case) to a secondary document or report which, although analytic in nature, was the primary published output of the original data gathering exercise rather than the data itself. Once the search for all of the DCAI elements was complete, the spreadsheet performed a final tabulation, assigning the article a score on the DCAI continuum (Mooney & Newton, 2012).

Results

Broadly, the results of the DCAI application suggest that data citation is poorly practiced across the journals surveyed in the three academic indexes. A perfect score of 36, where each of the DCAI elements would be found in the references section of the article (x3), was obtained by none of the articles in our sample. There were very few explicit citations of digital research data found in the reference sections of our samples. Rather, where reference was made to these data at all, it was most specifically to the dataset title in the main body of the article.

Figure 3 (following page) represents a frequency distribution of the DCAI scores and shows that the majority of the papers are on the low end with scores ranging from two to eight. The score of two represents the common practice (41.5 percent of articles) of mentioning in-text basic identifying information about the data, in most cases provision of the dataset title. Those with scores ranging from four to eight (41.5 percent of articles) mainly represent articles that included a footnote with some explanatory information regarding the data or may have mentioned more than one element within the text (e.g., author). The remaining papers (17 percent of articles) fall in the upper range of 24 to 28. These reflect the appearance of data citations in the reference list of a paper, but the citation is missing elements such as an electronic retrieval location, persistent identifier, publisher, or material designator. Not a single article from the sample cited a dataset with a persistent identifier.

The aggregate counts for DCAI elements located across all 65 articles in the sample can be reviewed in Figure 4 (following page), which was purposefully assembled in a manner similar to the tabular view of the data prepared in Sieber and Trumbo (1995). Notable differences between Sieber & Trumbo's data and that presented here include

Figure 3. Distribution of DCAI score (n=65)

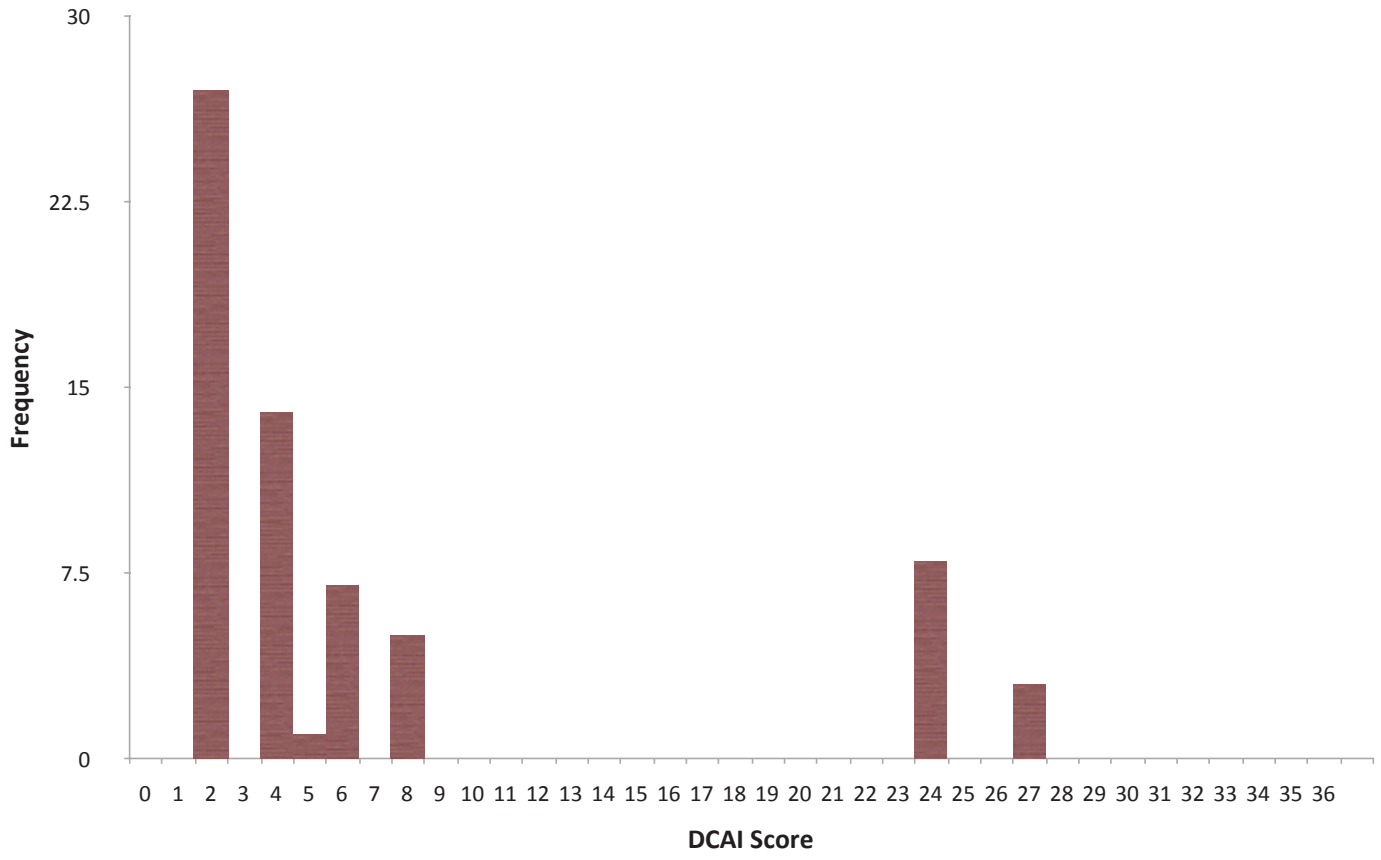


Figure 4. Counts of citation elements by location

Element and Code		Primary Citation Location and Location Code			
		[0] None	[1] Text	[2] Notes	[3] References
Author	[2]	40 (61.5)	14 (21.5)	0 (--.-)	11 (16.9)
Title	[2]	6 (09.2)	45 (69.2)	3 (04.6)	11 (16.9)
Date	[2]	54 (83.1)	0 (--.-)	0 (--.-)	11 (16.9)
Publisher	[2]	38 (58.5)	15 (23.1)	3 (04.6)	9 (13.8)
Material Designator	[1]	61 (93.9)	1 (01.5)	0 (--.-)	3 (04.6)
Electronic Retrieval Location	[2]	53 (81.5)	8 (12.3)	2 (03.1)	2 (03.1)
Persistent Identifier	[1]	65 (100.0)	0 (--.-)	0 (--.-)	0 (--.-)

both the slight variations on the elements themselves as well as the locations of the data citations in the text. For our study, we elected not to break out abstract as a separate location, for example, not only because the abstract functionally stands as a parallel text to the article but further because our sampling strategy involved a database query explicitly on the abstract. Regardless, the data in Figure 4 has been constructed to leverage the utility of comparison between the datasets. The data may be broken down further by looking at articles drawn from the separate databases (Figure 5, below).

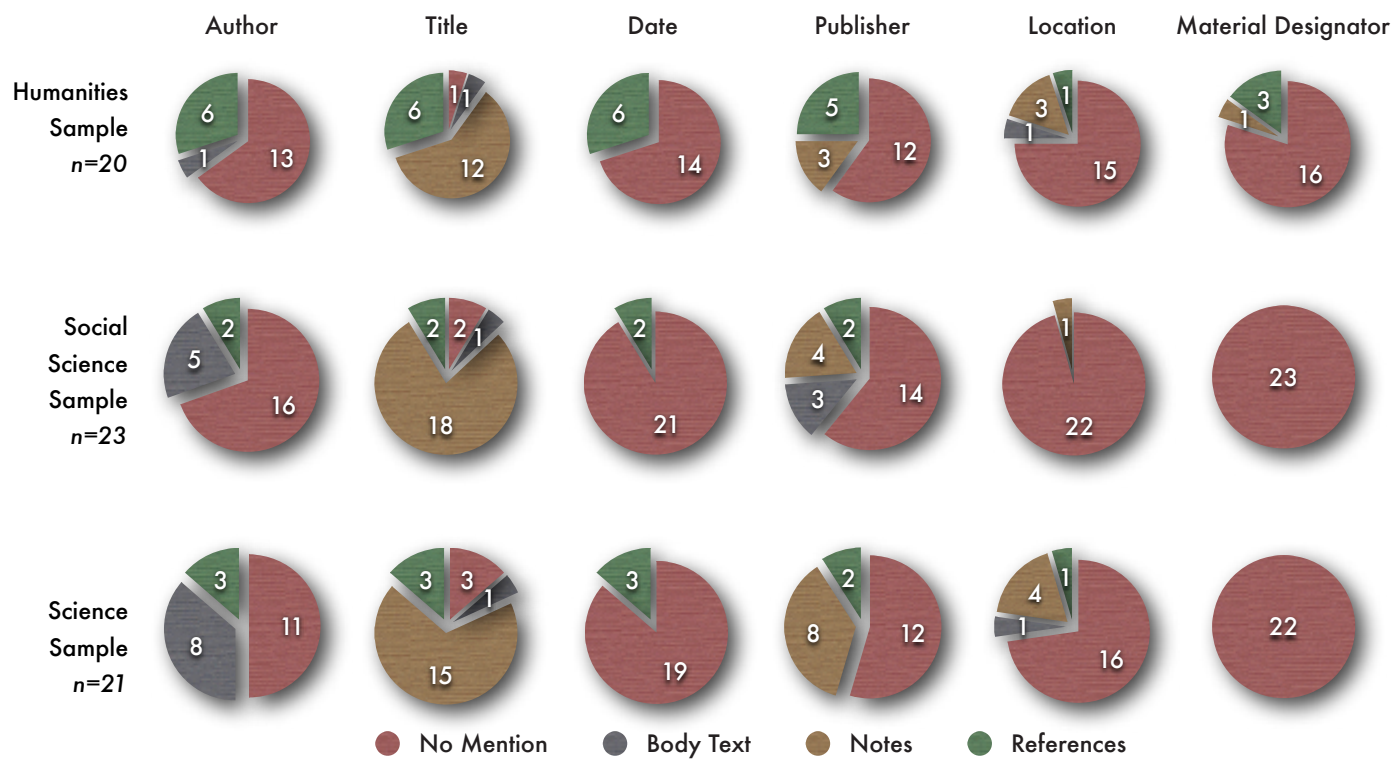
These data describe the sparse citation approach to the underlying digital research data within the articles in our sample. Each citation element registered more coded entries for none than any other category, excepting title, which was most frequently present in the text. Across all articles in our sample, therefore, textual references to the dataset title prevailed as the prominent mode of citation. Exempting coded references to none and the references to title coded as text, the next three greatest citation elements by volume were publisher in text (15), author in text (14), and a three-way split with author, title, and date in the references (11 each). Interestingly, citations including the date of publication of digital research data were found only in the references sections of the sampled papers and never in the body text or the notes. References to electronic retrieval location for digital research data were particularly few. Only 12 of the 65 datasets contained such references, and of these, only four were found as notations to the body text.

PART 3: DATA CITATION INSTRUCTIONS

Methods

The study sample of 65 articles yielded a journal sample of 44 individual periodical titles. For each of these titles, we reviewed the Author Guidelines/Instructions to Authors documentation to ascertain how citation formats are addressed and whether or not the citation of data is mentioned explicitly. Next, we followed up on the style manuals used by the journals and applied a modified DCAI to their treatment of data, with the greatest value placed on elements appearing in an example data citation. As data providers are another potential source of citation instruction, we attempted to identify the source location of each article's data in order to assess the form of

Figure 5. Placement of Dataset References within the Samples



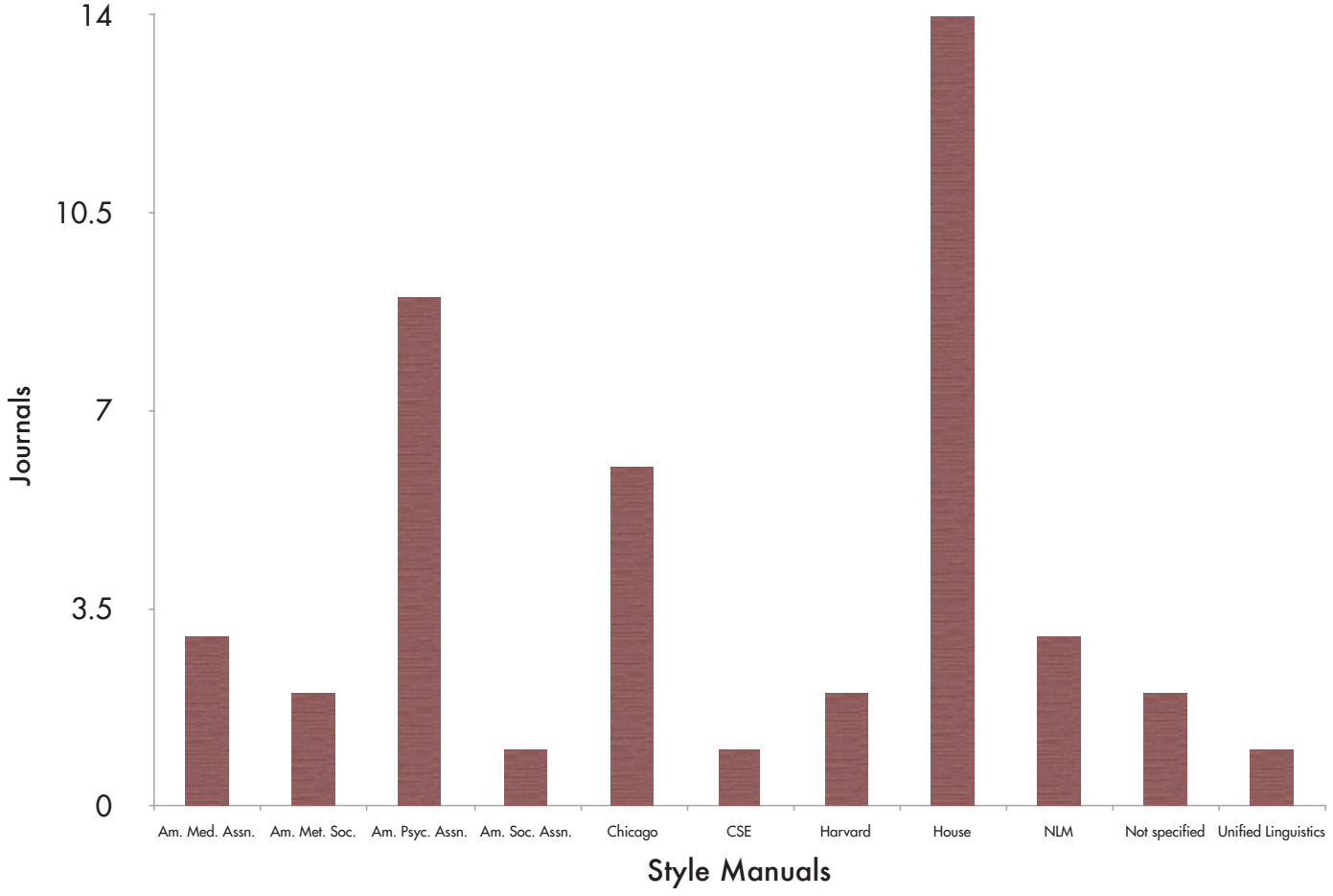
instructions for citation and reuse given there.

Journals approach the task of issuing guidelines for citation formatting in a number of ways. There are two main approaches: (a) a statement of referral to a style manual, or (b) provision of a house style sheet. A particular citation style may be specific to an individual journal, it may be reflective of multiple journals published by the same professional society or organization, or it may be in wide use across a discipline. Occasionally the distinction between a style manual and a house style sheet is blurry, given the multiplicity of individual journal formatting guidelines. The distinction made in this study for the qualification of a style manual is that (a) the citation style instructions exist as a distinct document outside of the discrete Author Guidelines documentation for the journal and (b) no direct referral is made to an outside style manual (although it may be stated that it is based on an outside style). Style manuals usually apply to multiple publications from a particular association, or are widely-used across a discipline. House style sheets generally apply to a single journal and provide specific instructions and examples without directing authors away to an outside source (although it may be stated that it is based on an outside style).

Results

Journal author instructions are largely silent on the issue of data citation. Figure 6 (below) illustrates the specific style manuals used by the sample journals. Given our criteria, 14 (32 percent) journals used a house style sheet, 28 (64 percent) referred to a style manual, and two (4 percent) did not specify any citation style at all. The journals using house style sheets are significant in that not one of these individual style sheets provide any instructions for the citation of data. A few (7 percent) journals include statements noting that statistical data should be cited, but a statistic is distinct from the main data under analysis within an article. However, there is some degree of attention paid towards data by journals, but instruction on the treatment of data within an article

Figure 6. Style Manuals Used by Journals (n=44)



is not necessarily considered a matter of citation. Some journals consider this a different matter of scholarly integrity and include policies on data availability for replication (e.g, *PNAS*, *Demography, Science*) or publishing supplemental data (e.g., *Sociology of Religion, Integrative and Comparative Biology*). In fact, seven (16 percent) journals (all in the sciences) specify that authors should deposit their primary data in domain repositories and include accession numbers in the text or notes of the article. Likewise, some style manuals discuss data sharing ethics separately from their treatment of reference formatting (e.g., AMA). These policies all relate to the treatment and availability of primary author produced data, rather than data used in secondary analysis.

For those journals that refer to a style manual, Figure 7 demonstrates the treatment of data within their citation formatting instructions. The issue of different definitions for data is germane as two main data types emerge: datasets and (non-bibliographic) databases. The style manuals that provide reference examples for datasets all do so with a slightly different terminology, but still

Figure 7. Data in Style Guides

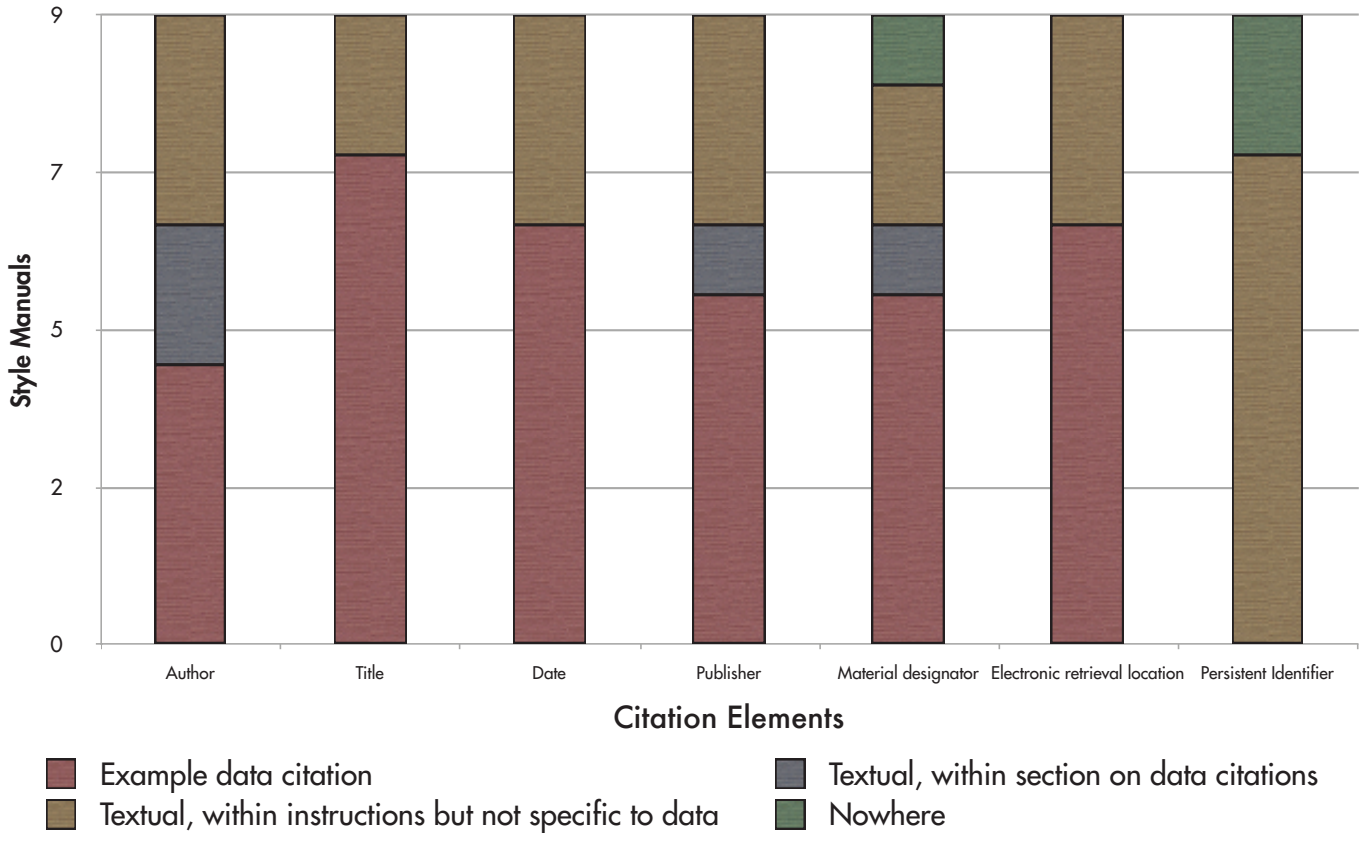
Style Manual	Data Type	Example Data Citation
AMA (Iverson et al, 2007, p. 70)	Databases	PDQ: NCI's Comprehensive Cancer Database. Bethesda, MD: National Cancer Institute; 1996. http://www.cancer.gov/cancerinfo/pdq/cancerdatabase. Updated December 18, 2001. Accessed April 29, 2004
AMS (American Meteorological Society, n.d., p. 6)	Digital media/NSIDC data	Jackson, T. J., and M. H. Cosh, 2003: SMEX02 watershed soil moisture data, Walnut Creek, Iowa. National Snow and Ice Data Center, Boulder, CO, digital media. [Available online at http://nsidc.org/data/nsidc-0143.html.]
APA (American Psychological Association, 2010, pp. 210-211)	Data Sets	Pew Hispanic Center. (2004). Changing channels and crisscrossing cultures: A survey of Latinos on the news media [Data file and code book]. Retrieved from http://pewhispanic.org/datasets/
ASA (American Sociological Association, 2010, p. 106)	Machine Readable Data Files	American Institute of Public Opinion. 1976. Gallup Public Opinion Poll #965 [MRDF]. Princeton, NJ: American Institute of Public Opinion [producer]. New Haven, CT: Roper Public Opinion Research Center, Yale University [distributor].
Chicago (University of Chicago Press, 2010, p. 764)	Scientific databases	GenBank (for RP11-322N14 BAC [accession number AC017046]; accessed October 6, 2009), http://www.ncbi.nlm.nih.gov/Genbank/.
CSE (Council of Science Editors, 2006, p. 558)	Databases on the Internet	IMGT/HLA Sequence Database [Internet]. Release 2.9.0 Cambridge (England): European Bioinformatics Institute. 2003 - [update 2005 Jun 1; cited 2005 Jun 22]. Available from http://www.ebi.ac.uk/imgt/hla/
Harvard	None	None
House	None	None
NLM (Patrias, 2007, Chapter 24)	Part of a Database on the Internet	Entrez Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [date unknown] - . Haloarcula marismortui ATCC 43049 plasmid pNG200, complete sequence; [cited 2007 Feb 27]. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome&cmd=Retrieve&dopt=Overview&list_uids=18013
Unified Linguistics	None	None

address the same format of data. Datasets are commonly associated with research in the social sciences, where survey data are often published as datasets deposited in data archives as individual discrete packages, or a complete set of data. Non-bibliographic databases are more often used in the sciences where data such as gene sequences or astronomical objects are recorded following a uniform format. A researcher may query a database to meet certain parameters and choose certain subsets of records (or an individual record) to analyze. Those style manuals including reference examples for datasets (APA, ASA, and AMS) are used by 12 (27 percent) of journals in our sample. This is similiar percentage to those style manuals that address databases (AMA, Chicago, CSE, and NLM) at 13 (30 percent), making a total of 25 (57 percent) journals that direct authors to style manuals that provide direction for the citation of some type of data.

Despite heterogeneity in data formats and definitions, the DCAI shows that there are common elements between and among citations for disparate data types. A modified DCAI was applied to the sample style manuals, with the greatest value placed on elements appearing in an example data citation. Figure 8 (below) illustrates the location of instructions within the style manuals (excluding the various house style sheets) for the citation elements in the DCAI. Seven of nine manuals address data specifically and the example data citations in all seven are equal only in the inclusion of the title. The majority of the remaining citation elements are addressed within the overall instructions but do not always appear within the example citation. While most of the style manuals discuss persistent identifiers in the context of journal article DOIs, they are not given in examples for data. This is likely to change in future style manual editions, as organizations like DataCite begin to assign unique identifiers to digital data.

Although it was impossible to track down the data providers for each article in our sample given the lack of complete citations, in roughly half (34) of the sample articles it was possible to identify a website providing access to the data. Out of this group, 68 percent (23) instructed data users to include a formal citation, with

Figure 8. Count of Citation Element by Instruction Location within Manuals (n=9)



56 percent (19) providing a sample formatted citation for inclusion in the reference list. This is promising, but there is still a lack of complete consistency as 26 percent (9) of data providers request an acknowledgment rather than a full citation (e.g., a footnote) and 15 percent (5) ask for a citation to a related literature references rather than the data set itself. (Note: Percents do not add up to 100 as three providers are counted twice for giving two separate options).

DISCUSSION

The first portion of the study elicited unsurprising results: even across the randomly selected cross-disciplinary sample, citation of digital research data is a rarefied activity. Where it is done, it is done rather completely, with enough attribution to perform a successful lookup of the cited datasets. The small sample, however, did not begin to suggest a groundswell of citation activity, and even in the most promising examples, sufficient citations with persistently identified online locations of these data were nonexistent. This may suggest a lack of community cohesion around the best practices for dataset citation or indeed around definitions for digital research data in the first place, although these shortfalls are rapidly being amended through initiatives such as DataCite and funder expectations on the management of research data, as can be seen evolving through agencies such as NSF. Regardless, the prevailing practice suggested in the sample remains a textual nod to the dataset title, with the work of identifying, acquiring, and verifying this data left as an exercise to the reader.

The citing behavior of authors is inextricably tied to the sources of instruction for academic and scholarly writing. Citation of traditional written works is a normal and expected behavior under-girding the foundation of science itself, passed down from generation to generation in the classroom (Kaplan, 1965). That the citation of data lags behind as a normative practice must be seen in the state of instructions provided to researchers from style manuals, journal policies, and data providers, along with their diffusion and enforcement. Despite approximately half of the journals pointing authors to style manuals that do address data citation, the majority of papers still failed to provide an adequate data citation. This shows a lack of awareness and understanding of data as a citable source on par with more traditional materials.

That data fails to conform to a uniform definition or type is a confounding factor. The style manuals in our sample either referred to datasets or non-bibliographic databases, but the articles in our sample used a wide assortment of data types ranging from anthropological field notes to satellite images, etc. It may be that authors do not readily recognize that their data has any association with the data reference type in a style manual. There is room for subject-specific style manuals to identify and refer to any data types in common use within disciplines and to include both datasets and non-bibliographic databases as prevailing high-level data types.

Requiring citations for both primary data in repositories and data used in secondary analyses would also allow for citation tracking metrics. Documenting reuse is a chief concern among data archives and is also important for individual data creators. Unfortunately, there is currently no systematic way to track data reuse (Piwowar, 2010; Schneider, 2006). Data providers can take steps to ensure citations by including robust metadata, suggested citations, and by making citation a condition of reuse.

Citation Styles

Citations must allow for access and provide acknowledgment in a standardized, concise, and intuitive fashion. One important outcome of the DCAI is the identification of common data citation elements. These are essentially the same elements needed to identify any type of resource; ergo published data used in secondary analysis studies are as citable as traditional formats. Specific standards for the citation of data and other resources will likely continue to evolve along with new ways of publishing, as with the recent introduction of persistent identifiers as an important piece of access information. Yet the upkeep of individual reference styles is far from being an insurmountable challenge, as these common citation elements remain unchangeable regardless of the specific source type in hand.

Although key citation elements are common among different source types, there are specific pieces of information of particular relevance to distinct formats. For example, volume and issue are pertinent to journal citations. Data as a format is special in that it can carry a two part publisher statement: both a producer and a distributor. It is common for data to be created and compiled at one institution and then disseminated via a third-party data archive. As a result, complete

bibliographic information about data can sometimes create lengthy citations. Conciseness is an important principle of citation and some styles tend more strongly towards efficiency and compactness in citations than others, but the provision of comprehensive information should be of greater value.

To this end, publisher statements should not be omitted when an electronic retrieval location is provided (see for example American Psychological Association, 2010, pp. 203, 211, which *does* endorse this practice). URLs and persistent identifiers are not always human readable in such a way that the publisher can be identified, and in the changing nature of the online environment they cannot perpetually be counted on to provide access. Even persistent identifiers can be prone to technical problems, as the GPO PURL server hardware failure in August 2009 illustrates. As publications are increasingly born digital and are no longer constrained by print era limits on space, value can accordingly be placed on completeness over compactness.

As the creation of the DCAI shows, there are a number of sources that style manuals, journals, and data providers can look to when setting their own standards. Additional sources of citation guidance can also be identified. Within a discipline one could look to those that are leading the way with established guidelines, especially widely used data archives. The Digital Curation Centre published a "how-to" guide that serves as a useful starting point (Ball & Duke, 2011). Style standards are always somewhat of a moving target, as new editions come out and source formats change over time. Despite the many valuable functions of style, it is secondary. Most important is an inclusion of sufficient bibliographic metadata elements in order to allow for access and acknowledgment.

CONCLUSION

Although the sample size of the present study is smaller when compared with the pioneering report of Sieber and Trumbo, it would appear that data citation remains an infrequent and haphazard activity another 15+ years in to the digital age. In both studies, for instance, in-text references to the dataset title account for the majority of citations, and references to the name of the data creators and publishers are scarce or not prominently featured. The state of play therefore appears not-yet-welcoming of the new era of data sharing in which researchers will want to rely on assurances of the assignment of credit when contributing digital research data to the scientific community.

There remain opportunities, however, for librarians, institutional repository managers, and library data curation specialists to affect positive change as information professionals operating as close collaborators with scientists, researchers, and other domain specialists by:

(1) Promoting the use of data citation standards (such as the one provided by the DataCite metadata initiative), which uses the publishing convention of the DOI to confer aspects of citability to digital research data while tying these data closer into the body of published literature for which regular, thorough attribution and citation of scholarly sources is formalized. Such promotion might begin through online library instruction (e.g., LibGuides), but would also be impactful in embedded librarian research contexts in addition to classroom bibliographic instruction opportunities and research and data consultations.

(2) Focusing conversations with the campus community on opportunities for data sharing—both on emerging conventions that facilitate such sharing (e.g., Creative Commons CC-Zero waivers) and on library- or university-provisioned repositories and archives that facilitate open sharing. Data citation and data sharing (as component parts of data publishing, data archiving, and data discovery) are evident in the prevailing models of the research data lifecycle (MIT Libraries, 2009). The work of librarians in scholarly communication support roles, promoting and developing services that reinforce the data lifecycle model among researchers, sends a strong message about the need for standardization and consistency in digital research data citation.

Shifting cultural norms is a slow process. Author practice and instruction is variable, but the foundation of citations as a hallmark of scholarly integrity makes the move towards the consistent citation of data likely. Data citation is deeply entrenched in the data-sharing movement. There is a bit of the chicken/egg dilemma here, as one clearly supports the other. Ultimately, the recognition of data as a significant contribution to the scholarly record is needed. Data publication is enabled by the utility of relatively recent digital technologies.

Coupled with forces such as the demand for efficient use of research dollars (by allowing for reuse of data), transparency, replication, and developing cultures of open access and sharing, data publication will gain recognition as a publication of record. Data citation is a necessary corollary to data publication. As the acceptance of data as a significant contribution to the scholarly record grows, data citation will become a mundane practice that fades into the background and is taken for granted.

REFERENCES

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, *13*(3/4). doi:10.1045/march2007-altman

American Meteorological Society. (n.d.). *Author reference/citation guide*. Retrieved from http://www.ametsoc.org/pubs/journals/author_reference_guide.pdf

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, D.C.: Author.

American Sociological Association. (2010). *Style guide* (4th ed.). Washington, D.C.: Author.

Ball, A., & Duke, M. (2011). *How to cite datasets and link to publications* (DCC How-to Guides). Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Baron, J. N. (1988). Data sharing as public good. *American Sociological Review*, *53*(1), vi-viii.

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, Mass: MIT Press.

Ceci, S. J. (1988). Scientists' attitudes toward data sharing. *Science, Technology, & Human Values*, *13*(1/2), 45-52.

Clubb, J. M., Austin, E. W., Geda, C. L., & Traugott, M. W. (1985). Sharing research data in the social sciences. In S. E. Fienberg, M. E. Martin, & M. L. Straf (Eds.), *Sharing research data* (pp. 39-88). Washington, D.C.: National Academy Press.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023-4038. doi:10.1098/rsta.2010.0165

Council of Science Editors. Style Manual Committee. (2006). *Scientific style and format: the CSE manual for authors, editors, and publishers* (7th ed.). Reston, VA: Council of Science Editors in cooperation with the Rockefeller University Press.

Dodd, S. A. (1979). Bibliographic references for numeric social science data files: Suggested guidelines. *Journal of the American Society for Information Science*, *30*(2), 77-82. doi:10.1002/asi.4630300203

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). *Sharing research data*. Washington, D.C: National Academy Press.

Green, T. (2009). *We need publishing standards for datasets and data tables* (OECD Publishing White Papers). OECD Publishing. doi:10.1787/787355886123

Hagge, J. (1997). Disciplinary style manuals as reliable guides to scientific discourse norms. *Technical Communication*, *44*(2), 129-141.

Heim, K. (1987). Social scientific information needs for numeric data. *Collection Management*, *9*(1), 1-53. doi:10.1300/J105v09n01_01

International Organization for Standardization. (2010). *ISO 690:2010. Information and documentation: Guidelines for bibiliographic references and citations to information resources*. Geneva, Switzerland: Author.

Iverson, et. al. (2007). *AMA manual of style: A guide for authors and editors* (10th ed.). New York: Oxford University Press.

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, *16*(3), 179–184. doi:10.1002/asi.5090160305

Kunze, J. A., Cruse, P., Hu, R., Abrams, S., Hastings, K., Mitchell, C., & Schiff, L. R. (2011). *Practices, trends, and recommendations in technical appendix usage for selected data-intensive disciplines*. UC Office of the President: California Digital Library. Retrieved from http://www.escholarship.org/uc/item/9jw4964t

LeClere, F. (2010, August 3). Too many researchers are reluctant to share their data. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Too-Many-Researchers-Are/123749

Marshall, E. (1990). Data sharing: A declining ethic? *Science*, *248*(4958), 952-957. doi:10.1126/science.2343306

MIT Libraries. (2009). Data life cycle: Data management and publishing. Retrieved from http://libraries.mit.edu/guides/subjects/data-management/cycle.html. Archived at http://www.webcitation.org/64UaPnEed on January 6, 2012.

Mooney, H. (2011). Citing data sources in the social sciences: Do authors do it? *Learned Publishing*, *24*(2), 99-108. doi:10.1087/20110204

Mooney, H. & Newton, M. P. (2012). Author behavior and

instructions for the citation of data (Academic Commons version) [data file and codebook]. New York, NY: Columbia University, Center for Digital Research and Scholarship. Retrieved from http://hdl.handle.net/10022/AC:P:13190

National Information Standards Organization, & American National Standards Insitute. (2005). *Bibliographic references*. Bethesda, MD: NISO Press.

Newton, M. P., Mooney, H., & Witt, M. (2010). *A description of data citation instructions in style guides*. Poster session presented at the 6th International Digital Curation Conference, Chicago, IL. Retrieved from http://docs.lib.purdue.edu/lib_research/121

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. *Eos Transactions, American Geophysical Union*, *91*(34), 297-298. doi:201010.1029/2010EO340001

Patrias, K. (2007). *Citing medicine: the NLM style guide for authors, editors, and publishers*. D. Wendling (Ed.). Bethesda, MD: National Library of Medicine.

Piwowar, H. (2010, November 9). Tracking dataset citations using common citation tracking tools doesn't work [Web log post]. *Research Remix*. Retrieved from http://researchremix.wordpress.com/2010/11/09/tracking-dataset-citations-using-common-citation-tracking-tools-doesnt-work/

Schneider, J. (2006, Spring). Why we need a data citation standard: Lessons learned from compiling ICPSR's Bibliography of Data-Related Literature. *ICPSR Bulletin*, *xxvi*(2), 9-12. Retrieved from http://www.icpsr.umich.edu/files/ICPSR/org/publications/bulletin/2006-Q1.pdf

Sieber, J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, *1*(1), 11-20. doi:10.1007/BF02628694

Stanley, B., & Stanley, M. (1988). Data sharing: The primary researcher's perspective. *Law and Human Behavior, 12*(2), 173-180. doi:10.1007/BF01073125

Starr, J., Ashton, J., Brase, J., Bracke, P., Gastl, A., Gillet, J., Heller, A., et al. (2011, July). DataCite metadata schema for the publication and citation of research data (Version 2.2). doi:10.5438/0005

Stodden, V. (2010). *The scientific method in practice: Reproducibility in the computational sciences* (MIT Sloan Research Paper No. 4773-10). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs*. Research Information Network. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., et al. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6). doi:10.1371/journal.pone.0021101

University of Chicago Press. (2010). *The Chicago manual of style* (16th ed.). Chicago, IL: Author.

Walker, J. R., & Taylor, T. W. (2006). Preface. *The Columbia guide to online style* (2nd ed., p. xvii-xxi). New York: Columbia University Press.

CORRESPONDING AUTHOR

Hailey Mooney
Data Services and Reference Librarian

Michigan State University Libraries
366 W. Circle Drive
East Lansing, MI 48824

mooneyh@mail.lib.msu.edu